DOCUMENT RESUME

ED 193 297 TM 800 612

AUTHOR Ekstrom, Ruth B.
TITLE Issues of Test Bias and Validity.
PUB DATE Sep 79
NOTE 17p.: Paper presented at the Annual Meeting of the American Psychological Association (New York, NY, September 3, 1979).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Criteria; Psychometrics; *Test Bias; Test Construction; *Testing Problems; Test Reliability; Test Selection; *Test Validity
IDENTIFIERS *Standards for Educational and Psychological Tests

ABSTRACT

Three areas of concern related to test bias and validity should be considered during the revision of the Standards for Educational and Psychological Tests. The first area concerns the sources and consequences of test bias. Five sources of bias have been identified: numerical bias, role bias, status bias, stereotypic bias, and familiarity bias. The revised Standards should address each of these types of bias. Another set of problems to be considered are those related to the choice of a criterion in validity studies. Such problems include evaluation of criteria in aptitude test development, an evaluation of predictors in selection procedures, and the appropriateness of techniques based upon person matching. The third main area of concern is that of construct validity and problems related to the use of factors as criteria. Psychologists should focus on the inferences that must be made in a selection situation, rather than on rigid validity categories, and then to plan an appropriate validation strategy. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

CN003011

Paper presented at the symposium on Revising the 1974 APA Test Standards,
American Psychological Association annual meeting, New York City
September 3, 1979

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. E Kstrom
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED193297

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Issues of Test Bias and Validity

Ruth B. Ekstrom
Educational Testing Service

One of the problems facing the Joint Review Committee on Test Standards is to develop materials and guidelines that will help both psychologists and the courts understand the common types and sources of test bias. In doing this, it will be especially important to demonstrate how test bias relates to test validity. This paper will discuss three main areas of concern related to test bias and validity: (1) sources and consequences of test bias; (2) the criterion problem in test validity; and (3) some of the special problems of construct validity.

Sources and Consequences of Test Bias. As psychologists, we have an obligation to see that the tests we develop, select, and use are as free as possible from all kinds of bias. To this end, it seems important that the revised Standards present an explanation of how tests may have bias in content, psychometric bias, or both, and alert test developers and users to the evidence for a relationship between content bias and psychometric bias. Thus, the Standards should describe and recommend ways to review tests for the several different types of content bias. These include: (1) numerical bias, (2) role bias, (3) status bias, (4) stereotypic bias, and (5) familiarity bias.

TM800612

- Numerical bias occurs when certain groups are predominant and overrepresented in a test content, while other groups are underrepresented or entirely missing. For example, Tittle, McCarthy, and Steckler (1974) have shown that women are rarely depicted in achievement tests; similar numerical bias occurs for other groups and in other types of tests.
- Role bias occurs when the members of a group are depicted only in a limited number of activities. Test content analysis has shown that, when women are portrayed in test items, they are shown primarily as housewives, teachers, and secretaries. This type of bias in test content carries the subtle message that all occupations are not open to women.
- Status bias, which is related to role bias, occurs when the members of a group are consistently shown in situations that do not involve authority or leadership. As an example, tests which depict school settings frequently show women as teachers but they rarely show women as principals or superintendents. Again, the subtle message of such a test is that women need not aspire to advancement up the career ladder.

- Stereotypic bias occurs when the members of a group are uniformly portrayed as having certain aptitudes, interests, or personality characteristics. The literature on bias in test content provide many examples of how women are portrayed as less able to deal with mechanical equipment, as always interested in homemaking and nurturant activities, as less assertive, etc. The stereotypic portrayal of women in tests, as in other material, has no substantiation in research on sex differences or in other factual evidence.

These kinds of content bias are not only an inaccurate portrayal of our culture and an indication of the differential expectations held for women and minorities, they also may have psychometric consequences and affect the scores of individuals exposed to such biased test items. In a recent research study, my colleagues and I (Donlon, Ekstrom, and Lockheed, 1979) found modest but significant correlations between an indicator of psychometric sex differences in response to achievement test items (percent males passing minus percent females passing the item) and sex bias in the content of these items; at three grade levels (grades 5, 8, and 10) females performed significantly better than males on items that contained more female references. We also found some evidence that role and status depictions were related to test performance. If test content affects the test scores of individuals, use of tests with biased content not only inappropriately portrays women and minorities but it also contributes to our underutilization of their talents.

SEP 2 9 1980

The fifth type of bias that I hope the developers of the new Standards will consider is what I call familiarity bias. It occurs when one group of individuals has a greater opportunity to experience the situations depicted in test items or to become acquainted with the kind of process presented by the item. An example occurred in an early version of a spatial visualization test for the Kit of Reference Tests for Cognitive Factors (French, Ekstrom, and Price, 1963) in which the subjects were told that the process involved in solving the problems is similar to "working with sheet metal." Such a statement probably biased this test in favor of males because it suggested that these items can only be solved by people who have some knowledge of sheet metal work. The identical process could have just as accurately been described as similar to "working with a dress pattern." I suspect that many mechanical aptitude and spatial visualization tests contain similar sex biased material and that this content may contribute to the findings of sex differences in certain aptitudes.

Therefore, I hope that the Committee will encourage a thoughtful review of the kinds of items included in all aptitude tests. Many aptitude tests, like our own Factor Kit, grew out of military supported research begun during World War II which usually involved all male and predominantly white populations. Out of this research, we developed certain traditions about the kinds of test items best used to measure a given aptitude. The time has now come for us to review these traditional item types, to determine if they are equally appropriate to the experiences of all individuals in our society, and to decide if there may be better or alternative ways of assessing some aptitudes.

The problem of familiarity bias is an especially complex one because it is possible for the same kind of learning to occur in different settings and as a result of different kinds of experiences. When different groups of individuals develop an aptitude through different kinds of experiences, it is important that tests of that aptitude sample from all of these activities, not just the activities done by one group. Research relating to the impact of experience on spatial abilities by McDaniel, Guay, Ball, and Kolleff (1978) found that substantially different patterns of activities and experiences by males and females were related to high and low spatial ability. Although both males and females with high spatial ability often participated in activities involving the use of a map or compass, high spatial male often repaired furniture, played pool, and used machine tools while high spatial females often sketched house plans, made jigsaw puzzles, built models, used hand tools, did weaving or macrame, and skied.

Differences in male and female test related experiences are sometimes concealed by the use of separate sex norms. This can create problems both in terms of planning individual futures and also when these sex-differences are interpreted as differences in the underlying aptitude rather than differences in the opportunity to learn. For example, separate sex norms are often found on mathematical aptitude tests. However, data collected by Educational Testing Service shows that most of the variance in male-female score differences on the SAT-M is attributable to differences in the number of years of high school mathematics taken by males and females. A similar study by De Wolfe (1977) found that most of the sex difference in high school juniors quantitative aptitude test scores was attributable to

differences in the number and kinds of mathematics courses taken by males and females. It would be most helpful for the Standards to include some guidelines to help test developers and users determine if and when separate norms are appropriate and the nature of the variables that should be used to differentiate between the norm groups. A discussion of how to deal with situations where race or sex is correlated with opportunity to learn might also be included as well as a discussion of how separate norms can limit or enhance opportunity.

Finally, I hope that the new Standards include a thorough discussion of the different models related to bias and fair selection, with special attention to the problem of internal inconsistency in many models. Psychologists need to have a clear understanding of the differences between these models and of the psychometric and value-related implications of using any given model.

The Criterion Problem and Test Validity. Another set of problems which the new Standards should probably consider are those related to the choice of a criterion in validity studies. This topic has received fairly extensive attention elsewhere (Gulliksen, 1976; Wild & Dwyer, 1977) so I shall only touch on it briefly here.

As Gulliksen has pointed out, test developers in the past tended to accept whatever criteria were available as final and then to create and revise predictor tests to agree with these criteria. He provides a number of examples of situations where high validity coefficients were obtained because of faulty criteria and where low validity coefficients occurred on tests that should have been valid if the correct criteria

had been used. Gulliksen states "a serious problem which should concern... investigators of test bias is the definite bias that exists in many criteria. ...The use of a properly designed and properly analyzed and properly interpreted battery of predictor tests to differentiate between biased and unbiased criteria is an important one that should receive increased emphasis from test developers." Gulliksen recommends that all aptitude test development include some evaluation of the criteria; I hope that the revised Standards will also include this recommendation.

In a related paper on sex bias in selection, Wild and Dwyer point out that the issue of bias in predictor variables has not been addressed by the various widely used selection models because they have tended to assume that the criterion is a valid, reliable and unbiased measure of performance for both males and females. As these authors state, there is evidence, at least in terms of grades and grade point averages that this assumption may be unwarranted; they then go on to provide evidence that "grade point averages (GPA) are not composed of equivalent units, that they have different meaning for different individuals, that GPA may be biased for some groups, that the GPA reliability is less than the reliability of the predictors, and that the GPA may not be a sufficient criterion." The Wild and Dwyer process for evaluating selection procedures, which includes an evaluation of both predictors and criteria, is one which I believe should receive thoughtful attention. One advantage of the Wild and Dwyer model is that it differentiates between ultimate and intermediate criteria, the latter being only approximate indicators of the former. Even if the ultimate criterion

is unbiased, intermediate criteria may exhibit bias or differential reliability. Bias can also occur in more novel criteria, such as those proposed as alternatives to grade point averages (Dunkle and Sandle , 1974). For example, instead of using grades, some employers consider part-time or summer employment as an indicator of potential in a job area. However, we live in an unperfect world where who you know rather than what you can do is at least sometime the basis of summer or part-time job hiring. Hence, the use of summer jobs as a criterion would probably result in racial and social class bias for job selection. Also, young women are more frequently given clerical jobs as part-time or summer work than are young men. Using this kind of job experience as an indicator of job aptitude would almost definitely result in sex biased job selection for women.

Another problem related to criterion choice and test validity involves the appropriateness of techniques based on person matching, such as those advocated by Kuder (1979). It is my belief that this technique cannot be entirely bias free and that its use tends to perpetuate past race, sex, and class discrimination (see Tenopyr's discussant paper for a reference to research which supports this view) because it may rely on variables that are not directly related to successful achievement but which are correlated with race, sex, socioeconomic background, experience, and other personal characteristics. In short, I think that there are better techniques to select management trainees than to find out if an individual enjoys playing golf or polo.

Special Problems of Construct Validity. Because construct validity involves a theoretical variable rather than a directly observable behavior, the problem of an appropriate criterion is especially difficult. In most cases we have little idea of how to define and measure the ultimate criterion for constructs such as logical reasoning ability or aggression. While research on the localization of brain functions and on information processing hold the promise of helping us better understand the theoretical constructs we use, we must cope with today's construct validity problems within the current state-of-the-art in psychology. We need to remember that the pioneering work of Thurstone, Guilford, and others was done at a time when psychologists knew less about the functioning of the human brain and before much of the research on human information processing. Those of us who develop and use construct based tests have a responsibility to examine all aspects of contemporary psychological research to see what new insights and implications it has for our work.

The typical procedure in construct validity has been to develop or examine groups of tests or test items which appear to be measuring the same thing (frequently these are tests or items that load on the same factor) and then to identify the psychological process which is common to the test or items. In some cases these tests or items are further validated against an external criterion, such as judges' ratings of certain behaviors.

There are several problems with this approach. First of all, the construct identification is all too often specific to a given test and how a particular psychologist interprets a theoretical variable. In our

review of the literature for the development of the Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, and Harman, 1976), it was not surprising to find that tests labeled with the same construct term (e.g., spatial ability) were measuring different things (e.g., visualization or perceptual speed). In fact, it was this kind of confusion that led to the development of the Factor Kits. In some cases, the hypothesized variable may not even be the real underlying cause of the observed relationship.

There are other problems related to the use of factors as criteria in construct validity. We know that the factor structure of a test can vary with the characteristics of the sample of subjects. We also know that thresholds for attribution can affect factor structure (Fiske and Bourne, 1977). This latter problem is especially important in relationship to personality factors. Not only are such factors a function of the instructions and the conditions under which the measurement takes place, but they are also subject to beliefs about the behaviors of males and females. Research has already shown that "mental health" is defined differently for males and for females and that these items on the MMPI are sensitive to the effect of race (Costello, 1977) because nonpathological racial variance contributes to elevated scores on some scales. When external criteria, such as judges' ratings are used to support the validity of personality scales, the possibility of bias becomes even greater. For example, the identical behavior in a male and a female may be rated as highly aggressive for the female but only moderately aggressive for the male. The research literature abounds with references about how identical products receive differential evaluation if they are thought to have been produced by a male or by a female.

All of these lead me to be concerned that the Standards provide us with some guidelines for examining theoretical constructs with suggestions for determining when a construct seems adequately identified if the constructs are the same across studies and across populations. In our Factor Kit work, we decided that a factor was adequately established if, at a minimum, it had been found in at least three different research studies conducted by at least two different researchers or laboratories. One approach to compare construct validity across groups is maximum likelihood confirmatory factor analysis. This technique was used by Rock and Werts (1978) to investigate the construct validity of the SAT by examining for invariance across different racial populations in: (1) similarity of factor patterns; (2) equivalent units of measurements; (3) standard errors of measurement of each factor; (4) reliability of each factor; and (5) a factor mean structure. This work showed that the SAT-V and SAT-M measure the same things in the same units with equivalent accuracy for each group. I hope that the Committee will recommend that this or other similar techniques receive wide usage when it is important to examine the goodness of fit of models based on theoretical constructs.

We also need to develop better ways to determine if the same test item measures the same thing in all groups of subjects. One can readily imagine situations where the identical item measures the ability to write basic English in one group of subjects and expressional fluency in another group.

Research on cognitive processes has shown that the identical test item may be solved by different methods or strategies (French, 1965; Gruen and Parkman, 1972; Sternberg, 1977). Specifically, French found that most individuals solved spatial visualization test problems by mental manipulation of the figure but that some individuals solved the problem with an analytical strategy which involved breaking the figure into components and examining them for similarities. Groen and Parkman report that most adults solve simple addition problems by memory look up but that some adults and most children solve these problems by incremental counting. This difference in strategies probably explains why some factor-analytic studies place such addition tests on a numerical facility factor while other studies find that they are part of a memory factor. Sternberg has reported four alternate forms of information processing used to solve logical syllogistic reasoning problems (spatial, linguistic, algorithmic, and mixed) and five patterns used in the solution of verbal analogies (synonym, antonym, category member, linear order, and functional). In short, identical test item content does not assure that identical information-processing strategies will be used by all test takers. The question remains as to whether or not we should consider that a test measures a single construct if it can be solved equally well by more than one strategy.

Carroll (1976) has analyzed a number of well-established cognitive factors and identified the operations and strategies required. There is, as yet, little research to show if some strategies are more effective than others on test items, like analogies or logical reasoning, which

appear to have multiple approaches or if, as Sternberg's work suggests, the ability to shift strategies is the key to success.

It seems likely that differences in test strategy arise from differences in the psychological availability of certain modes of thinking, perceiving, and remembering similar to what is involved in cognitive styles. These differences may be at least partially related to socialization and, if so, there may be some variation in preferred strategy from group to group. The research evidence does not seem to support differential test validity for different racial groups (Linn, 1978; Hunter and Schmidt, 1978) but there is some evidence that validities tend to be higher for females than for males.

It has been suggested that the Committee needs to provide test developers and other practitioners with guidance about the comparative strengths and weaknesses of content, criterion, and construct validity. An alternative approach, included in the draft of Division 14's new Principles for the Validation and Use of Personnel Selection Procedures, might be to encourage psychologists to focus on the inferences that must be made in a selection situation, rather than on these rigid validity categories, and then to plan an appropriate validation strategy.

If the Standards continue to differentiate between content, criteria, and construct validity, I hope they will give careful attention to the Eyde, Primoff, and Hardt (1979) paper from this symposium and provide guidance for differentiating between content and construct validity. It also seems especially important that the Standards warn about over-interpretation as a source of bias, such as, interpreting content referenced tests as construct-referenced. Entirely too much of the

psychological literature, in my view, describes an arithmetic achievement test as a measure of mathematical reasoning ability or a vocabulary test as a measure of general intellectual ability with little or no evidence to support these construct-like interpretations.

Despite the current state of confusion about construct validity and the need for much more research, I am confident that most of the constructs we use in our tests have sufficient psychological meaning to meet whatever scrutiny we subject them to. However, we may be in for a few surprises!

References

Carroll, J. B. Psychometric tests as cognitive tasks: A new 'structure of intellect." In L. B. Resnick (Ed.) The Nature of Intelligence. Hillside, N.J.: Erlbaum, 1976.

Costello, R. M. Construction and cross-validation of an MMPI Black-White scale. Journal of Personality Assessment, 1977, 41, 514-519.

De Wolfe, V. A. High school mathematics preparation and sex differences in quantitative abilities. EAC Report 77-7, University of Washington, June 1977.

Donlon, T., Ekstrom, R., & Lockheed, M. The consequences of sex bias in the content of major achievement test batteries. Measurement and Evaluation in Guidance, 1979, 11, 202-216.

Dunkle, M., & Sandler, B. Sex discrimination against students. Harvard University, Center for Law and Education, October 1974.

Ekstrom, R., French, J., & Harman, H. (Eds.). Kit of Factor Referenced Cognitive Tests. Princeton, N.J.: Educational Testing Service, 1976.

Eyde, L., Primoff, E., & Hardt, R. What should the content of content validity be? Paper presented at the Symposium on Revising the 1974 APA Test Standards, APA Annual Meeting, NYC, September 3, 1979.

Fiske, D. W., & Bourne, E. J. Thresholds for attribution can affect factorial structure. Educational and Psychological Measurement, 1977, 37, 713-723.

French, J., Ekstrom, R., & Price, L. (Eds.). Kit of Reference Tests for Cognitive Factors. Princeton, N.J.: Educational Testing Service, 1963.

French, J. The relationship of problem-solving styles to the factor composition of tests. Educational and Psychological Measurement, 1965, 25, 9-28.

Gruen, G. J., & Parkman, J. M. A chronometric analysis of simple addition. Psychological Review, 1972, 79, 329-343.

Gulliksen, H. When high validity may indicate a faculty criterion. Paper presented at the APA annual meeting, Washington, D.C. September 6, 1976.

Hunter, J. E., & Schmidt, F. L. Differential and single-group validity of employment tests by race: A critical analysis of three recent studies. Journal of Applied Psychology, 1978, 63, 1-11.

Kuder, F. Career matching, 1979. Address AMEG annual meeting, Las Vegas, Nevada, April 1979.

Linn, R. Single-group validity, differential validity, and differential prediction. Journal of Applied Psychology, 1978, 63, 507-512.

McDaniel, E., Guay, R., Ball, L., & Kolloff, M. A spatial experience questionnaire and some preliminary findings. Paper presented at the APA annual meeting, Toronto, August 1978.

Rock, D., & Werts, C. Construct validity of the SAT across populations -- an empirical confirmatory study. Princeton, N.J.: Educational Testing Service, December 1978.

Sternberg, R. Component processes in analogical reasoning. Psychological Review, 1977, 84, 353-378.

Tenopyr, M. Discussant remarks at the Symposium on Revising the 1974 APA Test Standards, APA Annual Meeting, NYC, September 3, 1979.

Tittle, C., McCarthy, K., & Steckler, J. Women and educational testing: Princeton, N.J.: Educational Testing Service, 1974.

Wild, C., & Dwyer, C. Sex bias in selection. Paper presented at the Third International Symposium on Educational Testing, Leyden, The Netherlands, June 1977.